

AEGIS-MIS: A Hybrid Explainable Misinformation Detection System Using Rule-Based Analysis and Machine Learning

Bright Duffour^{1, 2, ✉}, Alberta Ayitey²

1. Computer Science, University of the Potomac, Washington, USA

2. Computer Science, Loyola University Chicago, Chicago, USA

Received: May 17, 2026 | Review began: June 21, 2026 | Review ended: July 27, 2026 | Published: July 31, 2026

© Copyright 2026

This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Misinformation detection has become increasingly critical in modern digital ecosystems, yet many existing approaches rely on complex black box models that limit interpretability. This paper presents Automated Explainable Guard for Information Security-Misinformation Identification System (AEGIS-MIS), a hybrid explainable misinformation detection system that integrates rule-based linguistic analysis with lightweight machine learning techniques. The system employs Term Frequency-Inverse Document Frequency feature extraction and classical classifiers, including Logistic Regression and Support Vector Machines, combined with an explainability module that provides transparent reasoning for each prediction. The proposed framework is evaluated across both controlled synthetic data and a real-world benchmark derived from the LIAR dataset. While strong performance is observed in structured prototype settings, results on benchmark data demonstrate moderate but realistic performance, highlighting the inherent challenges of misinformation classification. Comparative experiments show that model selection and feature engineering improve performance consistency, with Support Vector Machines achieving stronger class-level detection and feature union approaches enhancing robustness. The system is deployed as a lightweight web application and application programming interface, demonstrating practical usability alongside interpretability. The findings emphasize the importance of combining rule-based reasoning with machine learning inference, positioning AEGIS-MIS as an extensible and transparent framework for real-world misinformation detection.

Categories: Explainable AI, Security and Privacy, Natural Language Processing (NLP)

Keywords: explainable ai, misinformation detection, machine learning, natural language processing, cybersecurity, fake news detection, text classification, trustworthy ai

Introduction

The modern information environment is shaped by unprecedented speed, scale, and accessibility. Digital platforms have empowered individuals to publish, share, and consume information at rates that were previously unimaginable. While this has created opportunities for open communication and global exchange of knowledge, it has also created an environment in which misinformation can proliferate rapidly and with significant consequences. False or manipulative claims can influence public behavior, distort civic discourse, and undermine confidence in institutions. In sectors such as public health, politics, education, and security, the consequences of widespread misinformation are particularly serious [1].

How to cite this article:

Duffour B, Ayitey A (July 31, 2026) AEGIS-MIS: A Hybrid Explainable Misinformation Detection System Using Rule-Based Analysis and Machine Learning. Cureus J Comput Sci 3 : es44389-026-00195-8. DOI https://doi.org/10.7759/s44389-026-00195-8

The challenge of misinformation is not limited to factual inaccuracies. Many harmful narratives are persuasive not because they are obviously false, but because they are framed in ways that exploit uncertainty, fear, distrust, or emotional vulnerability. This makes automated detection difficult. Systems must often do more than identify incorrect statements. They must also detect linguistic patterns associated with manipulative narratives, conspiracy framing, or emotionally charged deception. As a result, the problem of misinformation detection lies at the intersection of natural language processing, machine learning, trust and safety, and information security [2].

Traditional rule-based approaches have long been used to identify deceptive or suspicious language patterns because their decisions are grounded in explicit logic and are therefore highly interpretable [3]. However, such systems may be brittle, overly rigid, or unable to generalize effectively when language evolves or misinformation narratives change over time.

To evaluate misinformation detection systems under realistic conditions, several benchmark datasets have been developed. Among the most widely used is the LIAR dataset, which provides short political statements labeled across multiple truthfulness categories and serves as a standard benchmark for misinformation classification research [4].

The increasing adoption of artificial intelligence (AI) in decision-making systems has also raised concerns regarding transparency and trustworthiness. Many modern machine learning models function as opaque systems whose internal reasoning is difficult for users to understand. Consequently, explainable AI techniques have emerged to provide human-understandable explanations for model predictions and improve user trust in automated systems [5].

Recent advances in transformer-based natural language processing have significantly improved text classification performance. Attention-based architectures have enabled models to capture contextual relationships within language more effectively than traditional approaches [6]. Building upon these advances, Bidirectional Encoder Representations from Transformers (BERT)-based approaches have demonstrated strong performance in misinformation detection tasks by leveraging contextual language representations and transfer learning techniques [7].

In addition to deep learning methods, lightweight machine learning frameworks remain widely used for practical text classification applications because of their efficiency, interpretability, and ease of deployment. Libraries such as Scikit-learn continue to provide robust tools for implementing classical machine learning pipelines in real-world environments [8].

A growing body of literature has further examined misinformation detection from both theoretical and practical perspectives. Comprehensive surveys have summarized existing detection methodologies, challenges, and future research opportunities in the field [9,10]. Benchmark datasets such as FEVER have also contributed to the advancement of automated fact verification and evidence-based misinformation analysis [11]. More recent studies have explored enhanced transformer-based approaches, including FakeBERT, which combines deep contextual representations with misinformation detection objectives [12].

Beyond predictive performance, researchers have increasingly emphasized the importance of interpretability in AI systems. Interpretable machine learning frameworks provide practical guidance for understanding and auditing model behavior in real-world applications [13]. Similarly, explainable AI research has introduced taxonomies, methodologies, and evaluation frameworks designed to improve transparency, accountability, and trust in AI systems [14]. Public datasets such as FakeNewsNet have further supported research by combining news content, social context, and dynamic information to facilitate more comprehensive misinformation studies [15].

This paper introduces Automated Explainable Guard for Information Security-Misinformation Identification System (AEGIS-MIS) as a hybrid response to these challenges. The system combines rule-based trigger detection with machine learning-based classification and merges the outputs through a hybrid scoring engine. In doing so, it attempts to preserve the transparency of symbolic reasoning while leveraging the adaptability of statistical learning. Beyond algorithmic design, the project emphasizes practical usability through deployment as a live web application and Application Programming Interface (API)-accessible service, demonstrating end-to-end functionality in a realistic operational environment.

How to cite this article:

Duffour B, Ayitey A (July 31, 2026) AEGIS-MIS: A Hybrid Explainable Misinformation Detection System Using Rule-Based Analysis and Machine Learning. *Cureus J Comput Sci* 3 : es44389-026-00195-8. DOI <https://doi.org/10.7759/s44389-026-00195-8>

The contribution of this work is threefold. First, it presents a lightweight but functional hybrid misinformation detection architecture that integrates deterministic and probabilistic reasoning. Second, it introduces an explainability layer that translates internal system outputs into analyst-readable reports. Third, it demonstrates that such a system can be packaged, deployed, and operated as a real application rather than remaining confined to offline experimentation.

Unlike many existing approaches that prioritize model complexity, this work emphasizes a balance between interpretability, deployability, and performance. The proposed system is intentionally lightweight, making it suitable for environments where computational resources, transparency, and rapid deployment are critical. Although the initial prototype was developed and validated using a controlled synthetic dataset, the present study extends the evaluation pipeline by incorporating a real-world misinformation benchmark derived from the LIAR dataset. This additional evaluation stage provides a more realistic assessment of the machine learning component under naturally occurring linguistic ambiguity while preserving the original objective of designing a lightweight and explainable system for real-world deployment scenarios.

Materials And Methods

System overview

AEGIS-MIS was designed as a modular misinformation detection framework consisting of several interconnected components, including a Flask-based web interface, a rule-based detection engine, a machine learning classification engine, a hybrid scoring module, an explainability engine, a logging mechanism, and a REST API endpoint. In addition, an offline training pipeline was developed to support model generation, evaluation, and retraining.

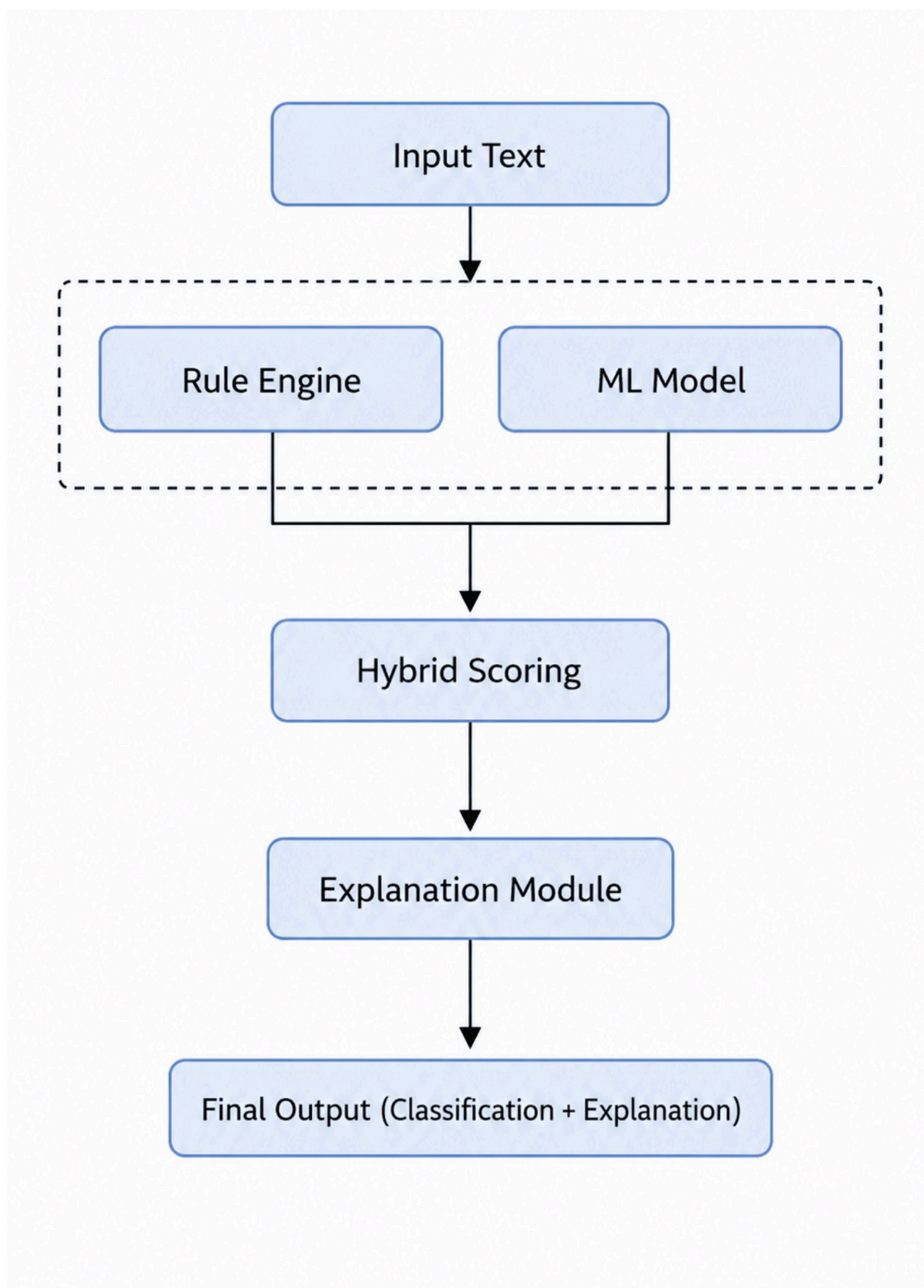
As illustrated in Figure 1, the system architecture follows a dual-path processing pipeline in which input text is analyzed simultaneously by rule-based and machine learning components before being merged into a unified scoring and explainability framework. Hybrid approaches that combine symbolic reasoning with statistical learning have been increasingly explored in misinformation detection because they provide improved interpretability while retaining predictive capability [1-3].

The online inference process begins when a user submits text through either the browser interface or the API. The submitted content is analyzed in parallel by the rule-based detector and the machine learning classifier. The rule-based component searches for predefined misinformation indicators, while the machine learning component evaluates linguistic patterns using a trained text classification model. The outputs of both components are subsequently combined through a hybrid scoring mechanism. The resulting score, confidence level, and supporting metadata are then passed to the explainability engine, which generates a structured report for the user.

This architecture reduces reliance on a single detection paradigm and allows explicit narrative indicators and statistical language signals to complement one another. The design further supports deployment scenarios requiring both machine-readable API outputs and human-readable analytical interpretation.

How to cite this article:

Duffour B, Ayitey A (July 31, 2026) AEGIS-MIS: A Hybrid Explainable Misinformation Detection System Using Rule-Based Analysis and Machine Learning. *Cureus J Comput Sci* 3 : es44389-026-00195-8. DOI <https://doi.org/10.7759/s44389-026-00195-8>



How to cite this article:

Duffour B, Ayitey A (July 31, 2026) AEGIS-MIS: A Hybrid Explainable Misinformation Detection System Using Rule-Based Analysis and Machine Learning. Cureus J Comput Sci 3 : es44389-026-00195-8. DOI <https://doi.org/10.7759/s44389-026-00195-8>

FIGURE 1: Architecture of the AEGIS-MIS hybrid misinformation detection system.

AEGIS-MIS, Automated Explainable Guard for Information Security-Misinformation Identification System; ML, Machine Learning

Rule-based detection

The rule-based detector was developed to identify linguistic markers frequently associated with manipulative, deceptive, or misleading narratives. The implementation relies on a curated collection of trigger phrases and patterns commonly observed in misinformation content, including references to hidden truths, secret cures, conspiracies, suppression of evidence, and claims of systemic manipulation [1,4].

Rather than assigning equal importance to all detected indicators, the system employs weighted scoring. Higher-risk indicators contribute more strongly to the final misinformation assessment than weaker indicators. For example, claims involving “secret cures” or deliberate information suppression receive greater weighting than isolated suspicious terms. This weighted strategy improves realism and provides a more nuanced representation of misinformation risk.

The rule-based component produces both a cumulative risk score and a list of matched trigger phrases. This dual output supports the explainability engine by allowing users to identify the specific narrative indicators that influenced the final assessment.

Machine learning-based classification

The machine learning component was implemented using a classical natural language processing pipeline consisting of Term Frequency-Inverse Document Frequency (TF-IDF) vectorization and supervised classification. TF-IDF is a well-established feature extraction technique for text classification that represents documents according to the statistical importance of terms while remaining computationally efficient for high-dimensional sparse text data [16,17].

Logistic Regression was selected as the primary baseline classifier because of its computational efficiency, interpretability, and demonstrated effectiveness for binary text classification problems [18]. During inference, incoming text is transformed using the same TF-IDF vectorization process applied during training. The classifier predicts whether the submitted content is more consistent with the reliable information or misinformation class and produces a probability-based confidence score that is incorporated into the hybrid scoring mechanism and explainability report.

For benchmark evaluation, the framework additionally evaluated a Linear Support Vector Machine (Linear SVM) and a feature-enhanced configuration combining word-level and character-level TF-IDF representations. These comparative models were included to assess the robustness of the proposed framework while maintaining lightweight computational requirements suitable for practical deployment [18].

All machine learning models were implemented using the Scikit-learn library [8].

For benchmark evaluation, the machine learning component was trained and evaluated using the LIAR dataset, a widely used benchmark for misinformation detection containing manually labeled political statements [4]. To support binary classification, the original six truthfulness categories were mapped into two classes. Statements labeled true, mostly-true, and half-true were grouped as reliable information, whereas statements labeled barely-true, false, and pants-fire were grouped as misinformation. The original LIAR training, validation, and testing partitions were preserved following this binary label conversion to maintain a standardized benchmark evaluation protocol.

Experimental setup

To evaluate the effectiveness of AEGIS-MIS, experiments were conducted using both a controlled synthetic dataset and a benchmark dataset derived from the LIAR corpus [4]. The synthetic dataset was used to validate the functionality of the hybrid architecture under controlled conditions, whereas the LIAR dataset provided a realistic benchmark containing naturally occurring misinformation examples.

How to cite this article:

Duffour B, Ayitey A (July 31, 2026) AEGIS-MIS: A Hybrid Explainable Misinformation Detection System Using Rule-Based Analysis and Machine Learning. *Cureus J Comput Sci* 3 : es44389-026-00195-8. DOI <https://doi.org/10.7759/s44389-026-00195-8>

Text preprocessing included normalization, conversion to lowercase, punctuation removal, and TF-IDF vectorization. For benchmark evaluation, the original LIAR training, validation, and testing partitions were preserved following binary label conversion. The resulting benchmark dataset consisted of 10,240 training samples, 1,267 validation samples, and 1,284 testing samples. The training partition contained 5,752 reliable-information samples and 4,488 misinformation samples. The validation partition contained 714 reliable-information samples and 553 misinformation samples, while the testing partition contained 668 reliable-information samples and 616 misinformation samples.

Four lightweight model configurations were evaluated: (1) TF-IDF with Logistic Regression (baseline), (2) an improved Logistic Regression model using refined feature preprocessing, (3) a Linear SVM, and (4) a feature-enhanced Linear SVM combining word-level and character-level TF-IDF representations. Logistic Regression and Linear SVM were selected because of their effectiveness for high-dimensional sparse text classification, while the feature-enhanced configuration was included to evaluate whether combining multiple textual feature representations could improve classification robustness without substantially increasing computational complexity.

Performance was evaluated using Accuracy, Precision, Recall, and F1 Score, which are widely accepted metrics for misinformation detection and text classification research [9,10]. These measures provide complementary perspectives on classification effectiveness, particularly in scenarios where false positives and false negatives carry practical consequences.

All experiments were implemented in Python using the Scikit-learn machine learning library [8]. The hybrid scoring mechanism combined rule-based detection outputs with machine learning predictions to generate final misinformation risk assessments and analyst-readable explainability reports.

Hybrid scoring mechanism

A central design objective of AEGIS-MIS was to integrate rule-based analysis with machine-learning predictions within a unified and interpretable scoring framework. The rule-based detector first computes an initial misinformation score by identifying predefined linguistic trigger phrases associated with deceptive or manipulative narratives. Each trigger phrase contributes a weighted value reflecting its relative importance. Lower-weight indicators such as fake, rigged, and deceived contribute one point, intermediate indicators including conspiracy, cover up, hidden truth, and they are lying contribute two points, while stronger indicators such as secret cure and they don't want you to know contribute three points. The cumulative rule-based score is obtained by summing the weights of all matched trigger phrases.

The machine-learning component independently evaluates the submitted text using the trained classifier and estimates the probability that the text belongs to the misinformation class. When the classifier predicts misinformation or produces a misinformation probability of at least 0.45, the system records an AI_MODEL_FLAG indicator and adds two points to the cumulative rule-based score.

The final misinformation score is therefore computed as:

$$S_{\text{final}} = S_{\text{rule}} + S_{\text{ML}}$$

where S_{rule} represents the cumulative weighted score produced by the rule-based detector and

$$S_{\text{ML}} = \begin{cases} 2, & \text{if the classifier predicts misinformation or } p_{\text{misinformation}} \geq 0.45 \\ 0, & \text{otherwise.} \end{cases}$$

The resulting score is mapped to three risk categories to support analyst interpretation. Scores below 2 are classified as Low Risk, scores from 2 to 3 are classified as Medium Risk, and scores of 4 or greater are classified as High Risk. This additive scoring strategy preserves the transparency of the rule-based component while allowing machine-learning evidence to strengthen the final assessment when statistical patterns indicative of misinformation are detected.

How to cite this article:

Duffour B, Ayitey A (July 31, 2026) AEGIS-MIS: A Hybrid Explainable Misinformation Detection System Using Rule-Based Analysis and Machine Learning. Cureus J Comput Sci 3 : es44389-026-00195-8. DOI <https://doi.org/10.7759/s44389-026-00195-8>

Explainability module

Explainability represents one of the distinguishing features of AEGIS-MIS. Rather than presenting only a numerical score or categorical label, the system generates a structured report containing the final risk level, misinformation score, detected trigger phrases, classification confidence, and an analyst-readable interpretation. An example of the explainability output generated by AEGIS-MIS is presented in Figure 2.

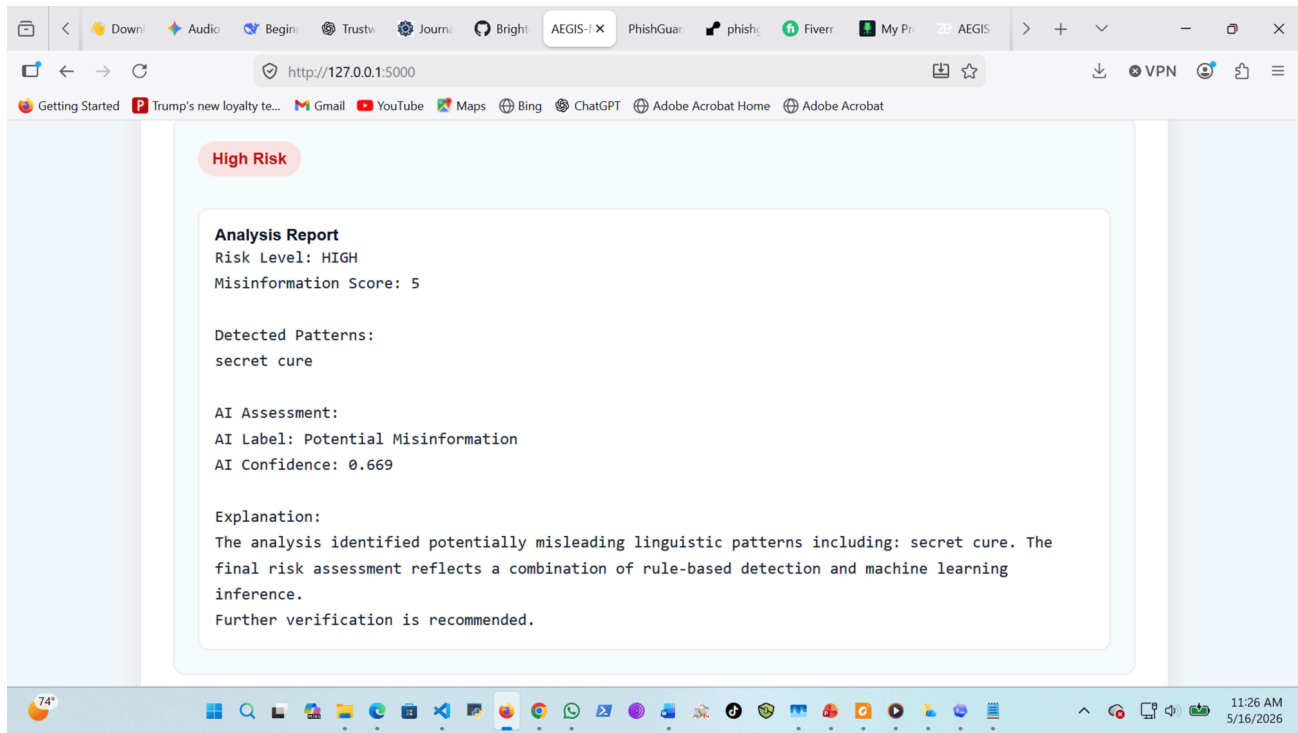


FIGURE 2: Example of explainability output generated by AEGIS-MIS.
AEGIS-MIS, Automated Explainable Guard for Information Security-Misinformation Identification System

The importance of explainability has been emphasized throughout the recent AI research because users frequently require justification for automated decisions before taking action [10-12]. By translating technical outputs into human-readable explanations, AEGIS-MIS improves transparency, user trust, and operational usability.

Deployment architecture

Beyond algorithmic design, AEGIS-MIS was implemented as a deployable service. The system operates as a Flask web application featuring both a browser-based user interface and a REST API for programmatic access. The live deployment demonstrates that the complete inference pipeline can function outside a local development environment and support practical use cases. The deployed interface is shown in Figure 3.

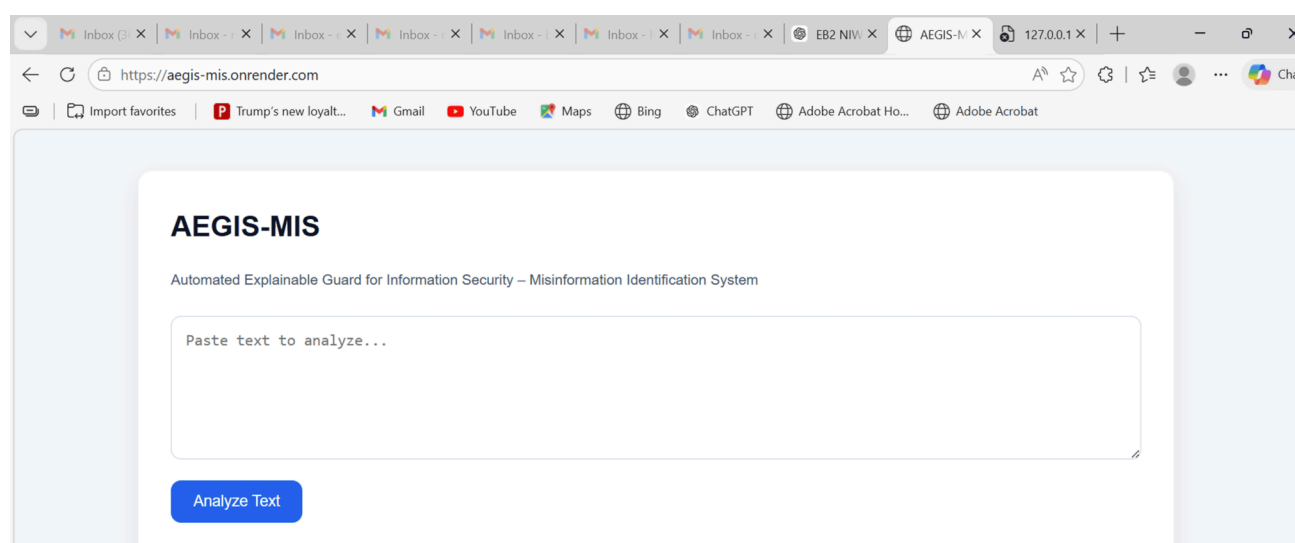


FIGURE 3: Web interface of the deployed AEGIS-MIS application.

AEGIS-MIS, Automated Explainable Guard for Information Security-Misinformation Identification System

Operational deployment is an important aspect of modern AI systems because practical impact depends not only on predictive performance but also on accessibility, reproducibility, and ease of integration into existing workflows [13]. The complete source code and deployment configuration for AEGIS-MIS are publicly available through the project repository, supporting transparency, and future development.

Results And Discussion

Evaluation and discussion

The evaluation of AEGIS-MIS was conducted in two distinct phases. During the first phase, the system was tested using a controlled synthetic dataset designed to validate the feasibility of the hybrid architecture, scoring mechanism, and explainability workflow. Under these controlled conditions, the classifier achieved strong predictive performance, confirming that the proposed architecture could successfully learn the intended misinformation patterns embedded within the prototype dataset. These results were useful for validating the functional design of the system and verifying integration between the rule-based, machine learning, and explainability components.

The second phase involved evaluation using a benchmark dataset derived from the LIAR corpus [7], which contains naturally occurring political statements exhibiting varying levels of truthfulness. On this benchmark dataset, the benchmark evaluation yielded an accuracy of 0.61, precision of 0.61, recall of 0.60, and an F1 score of 0.61 for the baseline TF-IDF and Logistic Regression model. Although these results are substantially lower than those obtained using the controlled synthetic dataset, they provide a more realistic assessment of misinformation detection performance under naturally occurring linguistic ambiguity.

The observed performance reduction is consistent with findings reported in previous misinformation detection research, which has demonstrated that naturally occurring misinformation often contains linguistic ambiguity, contextual dependencies, and subtle narrative framing that complicate classification [8-10]. Unlike controlled datasets, real-world misinformation frequently overlaps with legitimate opinion, satire, or partially true statements, making reliable classification significantly more challenging.

How to cite this article:

Duffour B, Ayitey A (July 31, 2026) AEGIS-MIS: A Hybrid Explainable Misinformation Detection System Using Rule-Based Analysis and Machine Learning. Cureus J Comput Sci 3 : es44389-026-00195-8. DOI <https://doi.org/10.7759/s44389-026-00195-8>

The benchmark evaluation strengthens the contribution of this work in two important ways. First, it demonstrates that the proposed architecture can operate on publicly available misinformation benchmarks rather than only on synthetic prototype data. Second, it highlights the limitations of relying exclusively on lightweight classical machine learning classifiers for misinformation detection. Rather than weakening the proposed framework, these findings position the current implementation as an interpretable and computationally efficient baseline that can be extended through future integration of more advanced language models.

Beyond predictive metrics, the explainability module remained one of the most practically valuable components of the system. The framework successfully identified linguistic trigger phrases, displayed machine learning confidence scores, and generated human-readable analytical interpretations. This aligns with recent explainable AI research emphasizing that transparency and interpretability are essential for building user trust in automated decision-support systems [11-13].

The deployment of AEGIS-MIS as a functional web application and REST API further distinguishes the project from purely experimental implementations. Even when benchmark classification performance remains moderate, the system demonstrates how rule-based reasoning, machine learning inference, and explainability mechanisms can be integrated into a lightweight end-to-end platform suitable for practical deployment. Consequently, the contribution of AEGIS-MIS should be viewed not as a claim of state-of-the-art predictive performance, but as an operational framework that balances interpretability, deployability, and analytical usefulness.

Although the benchmark performance is moderate compared with recent transformer-based approaches, the primary objective of AEGIS-MIS is to demonstrate a practical balance between interpretability, computational efficiency, and real-world deployability. This design philosophy makes AEGIS-MIS particularly suitable for operational environments where transparency, computational efficiency, and analyst trust are prioritized alongside predictive capability.

Model comparison

To further evaluate classifier performance, multiple lightweight machine learning models were tested using comparable feature representations. A baseline TF-IDF with Logistic Regression achieved an accuracy of approximately 0.61 on the benchmark dataset. Refinements to the feature extraction and classification pipeline produced modest improvements in overall performance while maintaining computational efficiency.

A Linear SVM model was subsequently evaluated. While its overall accuracy (0.61) was comparable to that of the baseline Logistic Regression model, it achieved the highest precision (0.65) and highest F1 score (0.64) among all evaluated configurations. These results indicate that the Linear SVM provided stronger discrimination of misinformation instances while maintaining comparable overall classification accuracy. Similar observations have been reported in previous text classification studies, where Linear SVM models frequently perform well on high-dimensional sparse text representations [18].

A feature-enhanced configuration combining word-level and character-level TF-IDF representations was also evaluated. This model achieved an accuracy of 0.62, precision of 0.62, recall of 0.62, and an F1 score of 0.62, demonstrating more balanced performance across evaluation metrics. Although its F1 score did not exceed that of the Linear SVM, the feature-enhanced approach produced more consistent overall performance.

Table 7 summarizes the comparative benchmark performance results obtained during experimentation.

How to cite this article:

Duffour B, Ayitey A (July 31, 2026) AEGIS-MIS: A Hybrid Explainable Misinformation Detection System Using Rule-Based Analysis and Machine Learning. *Cureus J Comput Sci* 3 : es44389-026-00195-8. DOI <https://doi.org/10.7759/s44389-026-00195-8>

Model	Accuracy	Precision	Recall	F1 Score
TF-IDF + Logistic Regression	0.61	0.61	0.6	0.61
Improved Logistic Regression	0.62	0.62	0.62	0.62
Linear SVM	0.61	0.65	0.63	0.64
Feature Union + SVM	0.62	0.62	0.62	0.62

TABLE 1: Benchmark performance comparison of the evaluated lightweight misinformation detection models.

SVM, Support Vector Machine; TF-IDF, Term Frequency-Inverse Document Frequency

Figure 4 presents the comparative performance of the evaluated misinformation detection models. The x-axis represents the evaluated model configurations, while the y-axis represents benchmark performance scores measured using Accuracy and F1 Score.

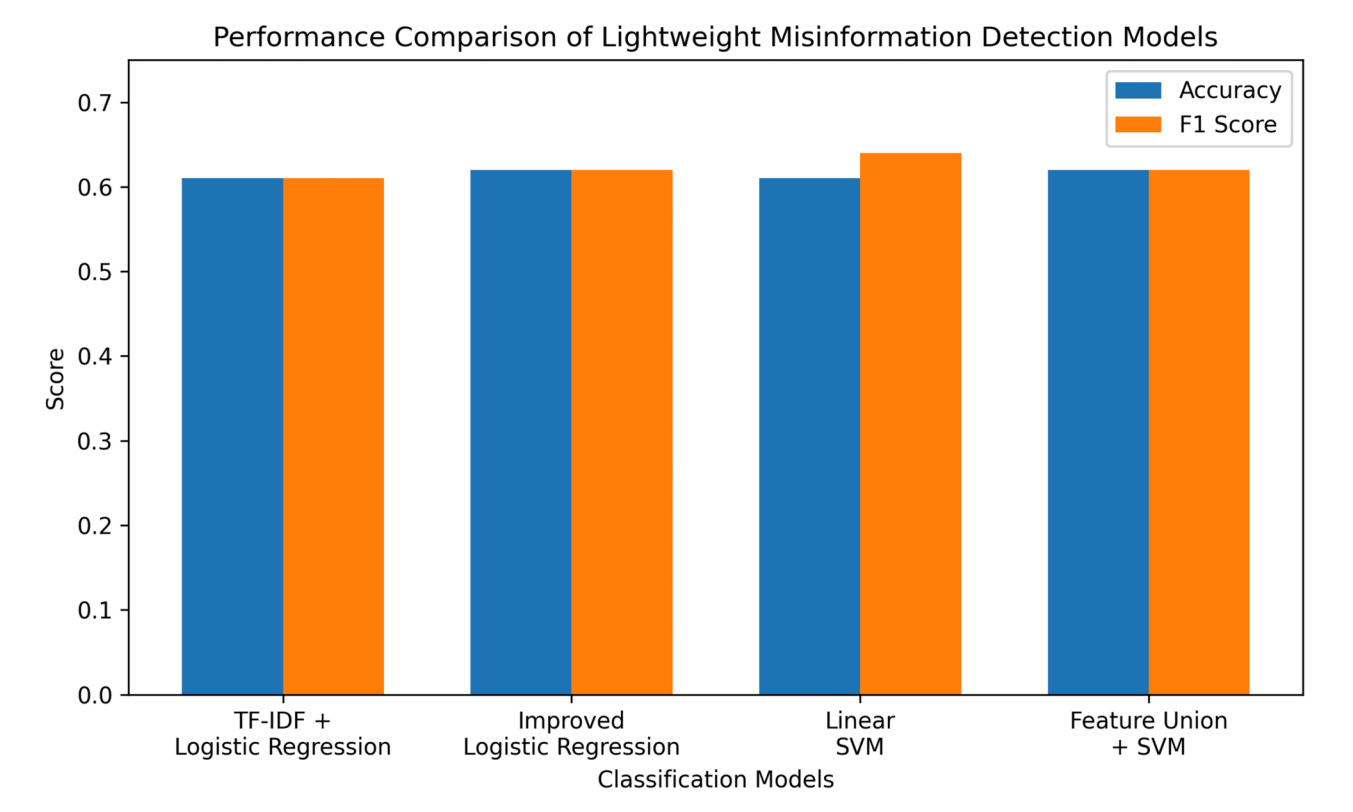


FIGURE 4: Performance comparison of lightweight misinformation detection models.

SVM, Support Vector Machine; TF-IDF, Term Frequency-Inverse Document Frequency

How to cite this article:

Duffour B, Ayitey A (July 31, 2026) AEGIS-MIS: A Hybrid Explainable Misinformation Detection System Using Rule-Based Analysis and Machine Learning. Cureus J Comput Sci 3 : es44389-026-00195-8. DOI <https://doi.org/10.7759/s44389-026-00195-8>

The comparative experiments clarify the rationale behind the hybrid design employed in AEGIS-MIS. The machine learning models demonstrated moderate success in capturing statistical language patterns associated with misinformation. However, performance remained constrained by the inherent ambiguity and contextual complexity of real-world language. In contrast, the rule-based component generated fully interpretable outputs through explicit identification of linguistic triggers, including exaggerated claims, urgency cues, and misleading narrative patterns.

While rule-based systems provide transparency, their effectiveness is inherently limited by the completeness of predefined rules and their inability to generalize beyond known patterns. Machine learning models address this limitation by learning broader statistical relationships from data but often sacrifice interpretability. The hybrid architecture combines these complementary strengths by leveraging machine learning for adaptability and rule-based analysis for transparency.

Following integration of hybrid scoring, the framework demonstrated improved sensitivity to linguistically subtle misinformation patterns. Inputs that did not contain explicit trigger phrases but exhibited statistical similarities to misinformation were more likely to be identified through the combined influence of machine learning confidence and rule-based analysis. The incorporation of classifier confidence into the scoring mechanism further improved differentiation of borderline cases by elevating them from low-risk to moderate-risk classifications when appropriate.

These findings suggest that the primary contribution of AEGIS-MIS lies not in maximizing benchmark accuracy, but in demonstrating how explainability, statistical inference, and operational deployment can be integrated into a unified misinformation detection framework. The results therefore support the feasibility of lightweight hybrid architectures as practical solutions for explainable misinformation detection while also highlighting opportunities for future improvement through advanced language modeling techniques [15].

Conclusions

This study presented AEGIS-MIS, a hybrid explainable misinformation detection framework that combines rule-based analysis with machine learning classification within a unified architecture. The system integrates trigger-based detection, statistical text classification, hybrid risk scoring, and explainability reporting to support transparent misinformation analysis. Experimental evaluation demonstrated strong performance in controlled environments and realistic performance on benchmark datasets derived from the LIAR corpus. The findings highlight both the challenges of real-world misinformation detection and the practical benefits of combining interpretable rule-based reasoning with machine learning inference. In addition to algorithmic evaluation, AEGIS-MIS was successfully implemented as a deployable web application and REST API service, demonstrating the feasibility of transforming explainable misinformation detection research into an operational platform. The results suggest that lightweight hybrid architectures can provide an effective balance between interpretability, deployability, and practical analytical utility for misinformation detection applications.

Future work will investigate transformer-based language models, including BERT, RoBERTa, and lightweight large language models, to evaluate whether improved predictive performance can be achieved while preserving the explainability objectives of the proposed framework. Future research will also incorporate additional publicly available misinformation benchmark datasets to improve generalizability and support broader comparative evaluation. Other directions include multilingual misinformation detection, adversarial robustness testing, adaptive rule generation, real-time social media monitoring, and human-in-the-loop verification mechanisms to further strengthen both the robustness and interpretability of the system.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

How to cite this article:

Duffour B, Ayitey A (July 31, 2026) AEGIS-MIS: A Hybrid Explainable Misinformation Detection System Using Rule-Based Analysis and Machine Learning. *Cureus J Comput Sci* 3 : es44389-026-00195-8. DOI <https://doi.org/10.7759/s44389-026-00195-8>

Concept and design: Bright Duffour, Alberta Ayitey

Acquisition, analysis, or interpretation of data: Bright Duffour, Alberta Ayitey

Drafting of the manuscript: Bright Duffour, Alberta Ayitey

Supervision: Bright Duffour

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The data (and/or code) supporting this study are openly available in Github at <https://github.com/Brightd4/AEGIS-MIS>, reference Brightd4/AEGIS-MIS.

References

1. Shu K, Sliva A, Wang S, Tang J, Liu H: [Fake news detection on social media: A data mining perspective](#). ACM SIGKDD Explorations Newsletter. 2017, 19:22-36. [10.1145/3137597.3137600](#)
2. Sharma K, Qian F, Jiang H, Ruchansky N, Zhang M, Liu Y: [Combating fake news: A survey on identification and mitigation techniques](#). ACM Transactions on Intelligent Systems and Technology. 2019, 10:1-42. [10.1145/3305260](#)
3. Ribeiro MT, Singh S, Guestrin C: ["Why should I trust you?": Explaining the predictions of any classifier](#). Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016, 1135-1144. [10.1145/2939672.2939778](#)
4. Wang WY: ["Liar, liar pants on fire": a new benchmark dataset for fake news detection](#). Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). 2017, 422-426. [10.18653/v1/P17-2067](#)
5. Doshi-Velez F, Kim B: [Towards a rigorous science of interpretable machine learning](#). arXiv. 2017, [10.48550/arXiv.1702.08608](#)
6. Vaswani A, Shazeer N, Parmar N, et al.: [Attention is all you need](#). arXiv. 2023, [10.48550/arXiv.1706.03762](#)
7. Devlin J, Chang MW, Lee K, Toutanova K: [BERT: Pre-training of deep bidirectional transformers for language understanding](#). Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019, 4171-4186. [10.18653/v1/N19-1423](#)
8. Pedregosa F, Varoquaux G, Gramfort A, et al.: [Scikit-learn: machine learning in Python](#). Journal of Machine Learning Research. 2011, 12:2825-2830.
9. Zhou X, Zafarani R: [A survey of fake news](#). ACM Computing Surveys. 2020, 53:1-40. [10.1145/3395046](#)
10. Bondielli A, Marcelloni F: [A survey on fake news and rumour detection techniques](#). Information Sciences. 2019, 497:38-55. [10.1016/j.ins.2019.05.035](#)
11. Thorne J, Vlachos A, Christodoulopoulos C, Mittal A: [FEVER: large-scale dataset for Fact Extraction and VERification](#). Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). 2018, 809-819. [10.18653/v1/N18-1074](#)
12. Kaliyar RK, Goswami A, Narang P: [FakeBERT: fake news detection in social media with a BERT-based deep learning approach](#). Multimedia Tools and Applications. 2021, 80:11765-11788. [10.1007/s11042-020-10183-2](#)
13. Molnar C: [Interpretable Machine Learning: A Guide for Making Black Box Models Explainable](#). (2022). <https://christophm.github.io/interpretable-ml-book/>.

How to cite this article:

Duffour B, Ayitey A (July 31, 2026) AEGIS-MIS: A Hybrid Explainable Misinformation Detection System Using Rule-Based Analysis and Machine Learning. Cureus J Comput Sci 3 : es44389-026-00195-8. DOI <https://doi.org/10.7759/s44389-026-00195-8>

14. Barredo AA, Díaz-Rodríguez N, Del Ser J, et al.: [Explainable artificial Intelligence \(XAI\): concepts, taxonomies, opportunities and challenges toward responsible AI](#). Information Fusion. 2020, 58:82-115. [10.1016/j.inffus.2019.12.012](#)
15. Shu K, Mahudeswaran D, Wang S, Lee D, Liu H: [FakeNewsNet: a data repository with news content, social context, and spatiotemporal information for studying fake news on social media](#). Big Data. 2020, 8:171-188. [10.1089/big.2020.0062](#)
16. Ramos J: [Using TF-IDF to determine word relevance in document queries](#). In: [Proceedings of the First Instructional Conference on Machine Learning](#). (2003). <https://www.scribd.com/document/339224468/TF-IDF-to-determine-word-relevance-in-document-queries-pdf>.
17. Hosmer DW, Lemeshow S, Sturdivant RX: [Applied Logistic Regression](#). John Wiley & Sons, Hoboken, NJ; 2013. [10.1002/9781118548387](#)
18. Cortes C, Vapnik V: [Support-vector networks](#). Machine Learning. 1995, 20:273-297. [10.1007/BF00994018](#)

How to cite this article:

Duffour B, Ayitey A (July 31, 2026) AEGIS-MIS: A Hybrid Explainable Misinformation Detection System Using Rule-Based Analysis and Machine Learning. Cureus J Comput Sci 3 : es44389-026-00195-8. DOI <https://doi.org/10.7759/s44389-026-00195-8>